CLAIMS

What is claimed is:

1. A method for mining a document containing dirty text comprising:

removing an instance of dirty text within said document to produce a cleaned document; and

performing a data mining operation on said cleaned document.

2. The method for mining a document containing dirty text as recited in Claim 1, wherein said removing further comprises replacing an instance of dirty text with a standard term.

3. The method for mining a document containing dirty text as recited in Claim 1, wherein said removing further comprises removing an instance of computer code from said document.

4. The method for mining a document containing dirty text as recited in Claim 1, wherein said removing further comprises removing a table from said document.

5. The method for mining a document containing dirty text as recited in Claim 1, wherein said performing a data mining operation further comprises identifying a sentence within said cleaned document by identifying a beginning and an end of said sentence.

6. The method for mining a document containing dirty text as recited in Claim 5, wherein said performing a data mining operation further comprises scoring and ranking said sentence.

7. The method for mining a document containing dirty text as recited in Claim 6, wherein scoring said sentence further comprises:

selecting scoring techniques operable for summarizing non-narrative, grammatically incorrect text;

selecting scoring techniques operable for summarizing narrative, grammatically correct text; and

using said scoring techniques to score said sentence.

8. The method for mining a document containing dirty text as recited in Claim 7, wherein said method further comprises generating a summary derived from said scored and ranked sentences.

9. The method for mining a document containing dirty text as recited in Claim 1, wherein said method further comprises selecting a text mining component based upon said data mining operation to be performed.

10. The method for mining a document containing dirty text as recited in Claim 1, wherein said method further comprises customizing said method by adjusting a parameter value.

11. A computer system comprising:

a bus;

a memory unit coupled to said bus; and

a processor coupled to said bus, said processor for executing a method for mining a document containing dirty text comprising:

removing an instance of dirty text within said document to produce a cleaned document; and

performing a data mining operation on said cleaned document.

12. The computer system as recited in Claim 11, wherein said removing further comprises replacing an instance of dirty text with a standard term.

13. The computer system as recited in Claim 11, wherein said removing further comprises removing an instance of computer code from said document.

14. The computer system as recited in Claim 11, wherein said removing further comprises removing a table from said document.

15. The computer system as recited in Claim 11, wherein said performing a data mining operation further comprises identifying a sentence within said cleaned document by identifying a beginning and an end of said sentence.

16. The computer system as recited in Claim 15, wherein said performing a data mining operation further comprises scoring and ranking said sentence.

17. The computer system as recited in Claim 16, wherein scoring said sentence further comprises:

selecting scoring techniques operable for summarizing non-narrative, grammatically incorrect text;

selecting scoring techniques operable for summarizing narrative, grammatically correct text; and

using said scoring techniques to score said sentence.

18. The computer system as recited in Claim 17, wherein said method further comprises generating a summary derived from said scored and ranked sentences.

19. The computer system as recited in Claim 11, wherein said method further comprises selecting a text mining component based upon said data mining operation to be performed.

20. The computer system as recited in Claim 11, wherein said method further comprises customizing said method by adjusting a parameter value.

21. A computer-usable medium having computer-readable program code embodied therein for causing a computer system to perform the steps of:

removing an instance of dirty text within said document to produce a cleaned document; and

performing a data mining operation on said cleaned document.

22. The computer-usable medium of Claim 21, wherein said removing further comprises replacing an instance of dirty text with a standard term.

23. The computer-usable medium recited in Claim 21, wherein said removing further comprises removing an instance of computer code from said document.

24. The computer-usable medium recited in Claim 21, wherein said removing further comprises removing a table from said document.

25. The computer-usable medium recited in Claim 21, wherein said performing a data mining operation further comprises identifying a sentence within said cleaned document by identifying a beginning and an end of said sentence.

26. The computer-usable medium recited in Claim 25, wherein said performing a data mining operation further comprises scoring and ranking said sentence.

27. The computer-usable medium recited in Claim 26, wherein scoring said sentence further comprises:

selecting scoring techniques operable for summarizing non-narrative, grammatically incorrect text;

selecting scoring techniques operable for summarizing narrative, grammatically correct text; and

using said scoring techniques to score said sentence.

28. The computer-usable medium recited in Claim 27, wherein said method further comprises generating a summary derived from said scored and ranked sentences.

29. The computer-usable medium as recited in Claim 21, wherein said method further comprises selecting a text mining component based upon said data mining operation to be performed.

30. The computer-usable medium as recited in Claim 21, wherein said method further comprises customizing said method by adjusting a parameter value.